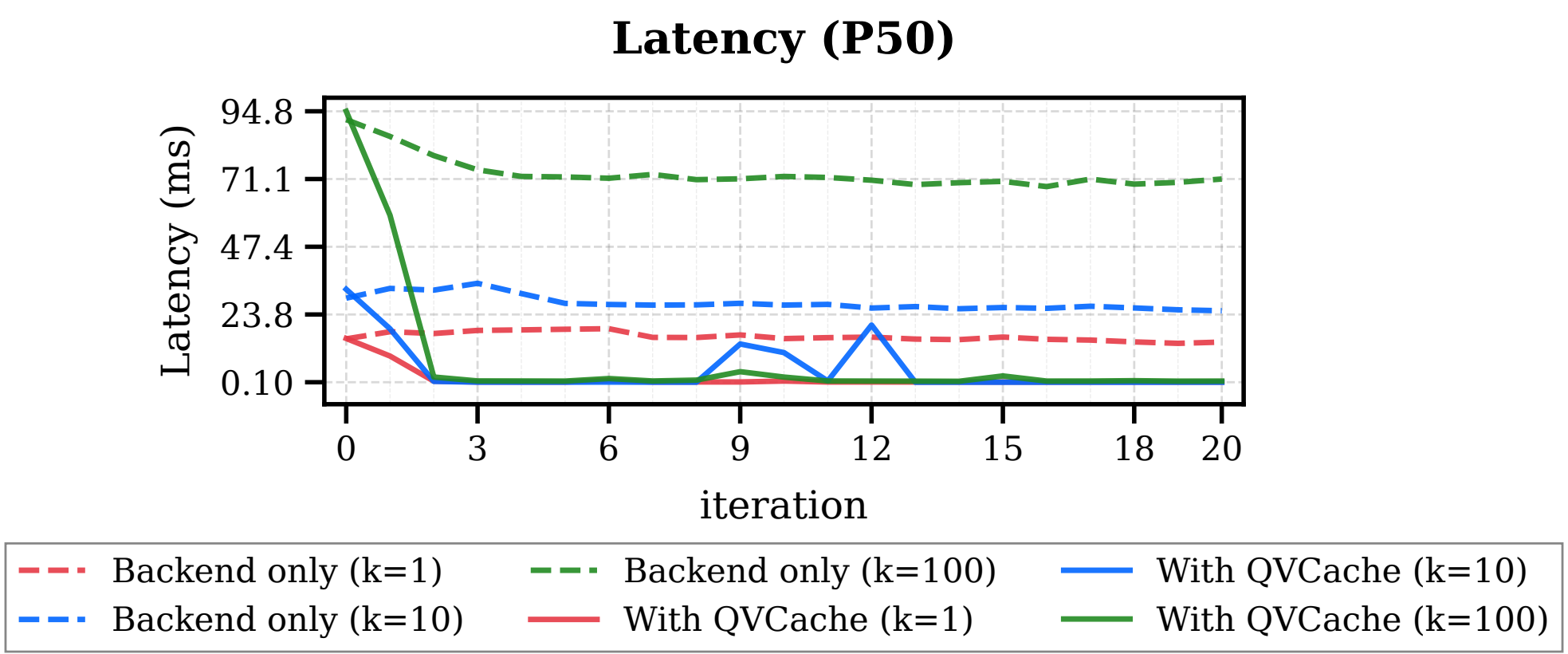
Latency (P50)
Latency (ms)
94.8
71.1
47.4
23.8
0.10
0
3
6
9
12
15
18
20
iteration
Backend only (k=1)
Backend only (k=10)
Backend only (k=100)
With QVCache (k=1)
With QVCache (k=10)
With QVCache (k=100)